# 20.9.18

1. Feature preparation
   1. Create a label.csv file with true locations of 1297 cells.
   2. Using DistMap create the label.csv file: the true cell location of the given 1297 cells in 3039 possible options.
   3. Changes 'na' to 'naa' and 'nan' to 'nana'.
   4. Moved in-situ genes to start of dge\_raw
   5. Simple scaling of data to [0,1].
2. Optimization: using pairs to select training and test samples.
   1. Train data includes 11300 random samples. Samples were built using cartesian product of bdtnp and dge files, where a label of 1 is added if the two lines are from a ‘matched’ location (using the label.csv file).
3. Test different models.
   1. Add dropouts.
   2. Use Softplus.
   3. Use Sigmoid in the last layer and and binary\_crossentropy as a loss function.
   4. Tried different architecture.
4. Check convergence using all 84 in situ genes.
   1. After 4000 epochs val\_acc stabilizes on 0.8838
5. Feature reduction: leave one out and Garson's methods.
   1. Use correlation test to validate feature reduction results.
   2. Leave one out seemed more efficient
   3. List of selected 60 genese below (in decreasing order).
6. Test with 60 genes.
   1. Reached val\_acc of 0.8885 after 2000 epochs.

# 11.10.18

1. Model changed to a ResNet like model.
2. Used Imputed\_data-magic\_dge.csv file.
3. Tried different dimension reduction technique using matrix factorization.
4. Summary of runs

|  |  |  |  |
| --- | --- | --- | --- |
| Number of genes | Model type | Avg. Validation accuracy | Remarks |
| 20 | Old model | 0.70 |  |
| 84 | New model | 0.91 |  |
| 20 | New model | 0.87 |  |

1.11.18

1. Found an error using model.fit: class\_weight is used only for loss calculation.

*history = model.fit(x=[X\_train], y=y\_train, batch\_size=5, epochs=20, verbose=1, validation\_split=0.2, callbacks=callbacks\_list,* ***class\_weight={0:0.55, 1:5.22}****)*

2. Changed main model from ANN to optimization of MAX(MCC)

- Select bests subset of genes that maximizes MAX(MCC) on correct ‘pairs’

- Almost no issues when downgrading to 60 genes

- Requires KMeans clustering when downgrading to 40 genes to get good results

- Hard to improve results when using 20 genes

3. Gene2Function works best when trained on in-situ genes (adult fly).

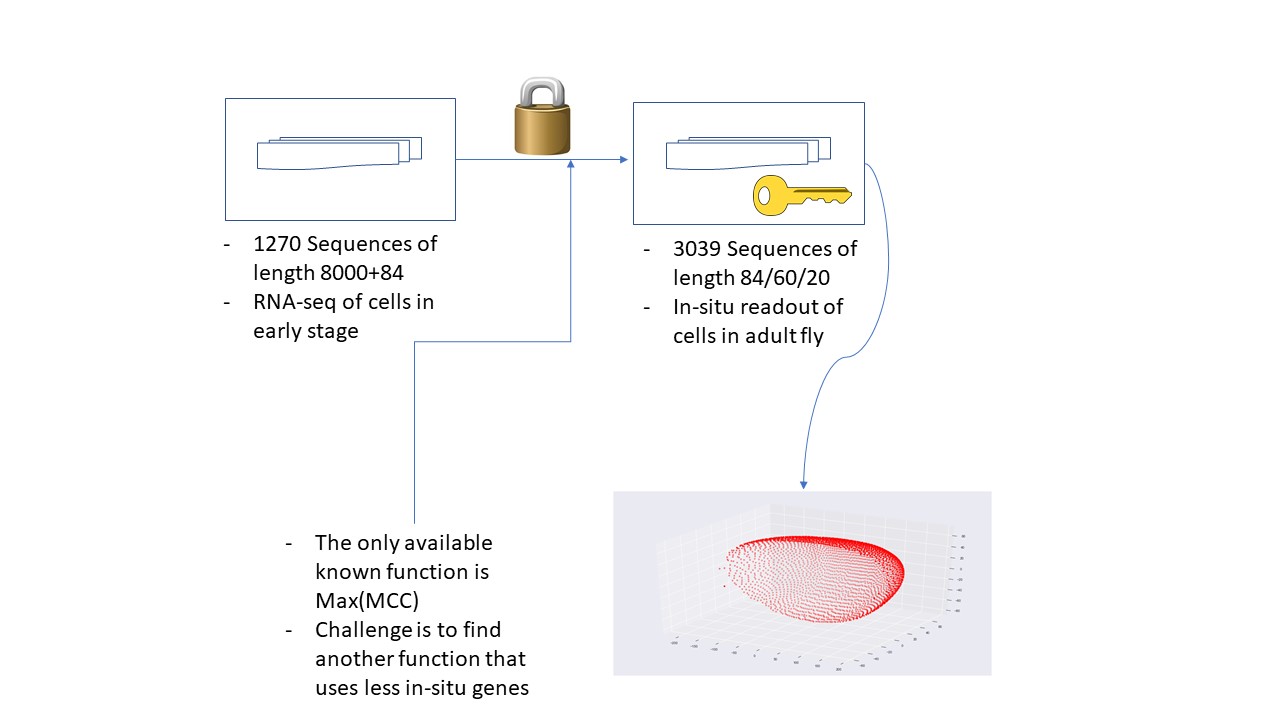
- Can predict geometrical location with R^2 > 0.97 (very very close to coordinate).

- But performs less than max(MCC) when train or tested on embryonic fly.

4. Next steps

- Search for relations between (dge) RNA sequences and Gene2Function sequences.

- Try the segment polarity genes wg, en, hh, ptc, ci.



# 18.11.18

1. In order to assist MAX-MCC (selecting top n candidates for a location) a new ANN based model is developed – predicting directly the BDTNP sequence (not a Boolean model as before that had low accuracy score).

- Main idea is to select cells that have the “same span”. Not just high (close) on MAX MCC but also on the same list of genes.

- Model is simpler (less parameters).

- Uses only the binarized versions of DGE and BDTNP.

- Using dropout and validation set to avoid overfitting.

2. Model performs well on 60 and 40 genes.

- In case of 60 genes sub-challenge we take n=8 highest MAXMCC values.

- In case of 40 genes sub-challenge we take n=2 highest MAXMCC values (and the rest are using the ANN model).

20-genes sub challenge:

3. Using ANN model we improved the previous results we had on the 20 genes sub-challenge (using only MAX MCC and some clustering approach). But still result is not satisfying (747 out of 1693 locations identified).

- In case of 20 genes MAXMCC is totally off and is not used for prediction, only the ANN model is used.

- Another approach: try using ANN models with 20 top candidates, and then use MAXMCC – but this proved non-useful as ANN prediction for a cell matches many locations in BDTNP (BDTNP has the same value for 10-20 rows). Obviously MAXMCC will be the same for all these rows.

- We need BDTNP rows to be as much different as possible on these 20 rows, in order to ‘assist’ ANN model to differentiate them.

- Starting from the full list of 84 genes (where our previously selected 20 genes are at the top) we run a program that uses different combinations and measure them (how many rows are different). This improved the result significantly to 899 out of 1693 locations.